

# SEARCH ENGINES

## INFORMATION RETRIEVAL IN PRACTICE

BOOK:

[HTTP://CIIR.CS.UMASS.EDU/DOWNLOADS/SEIRIP.PDF](http://ciir.cs.umass.edu/downloads/seirip.pdf)

SLIDES:

[HTTP://WWW.SEARCH-ENGINES-BOOK.COM/SLIDES/](http://www.search-engines-book.com/slides/)

All slides ©Addison Wesley, 2008

Slides adapted from Prof. W. Bruce Crof @UMASS

RETRIEVAL MODELS

SECTION 7.3

[HTTPS://CIIR.CS.UMASS.EDU/DOWNLO  
ADS/SEIRIP.PDF](https://ciir.cs.umass.edu/downloads/seirip.pdf)

All slides ©Addison Wesley, 2008

Information Retrieval in Practice

# Language Model

- *Unigram language model*
  - ▣ probability distribution over the words in a language
  - ▣ generation of text according to the probability distribution

e.g., 5 words in total {"girl", "cat", "love", "cute", "touched"}

A possible language model: [0.2, 0.1, 0.35, 0.25, 0.1]

# Language Model

- *Unigram language model*

- ▣ probability distribution over the words in a language
- ▣ generation of text according to the probability distribution

e.g., 5 words in total {"girl", "cat", "love", "cute", "touched"}

A possible language model: [0.2, 0.1, 0.35, 0.25, 0.1]

If we treat each document as a sequence of words, then the probabilities in the language model predict what the next word in the sequence will be.

# Language Model

- *Unigram language model*

- ▣ probability distribution over the words in a language
- ▣ generation of text according to the probability distribution

e.g., 5 words in total {"girl", "cat", "the", "cute", "touched"}

A possible language model: [0.2, 0.1, 0.35, 0.25, 0.1]

If we treat each document as a sequence of words, then the probabilities in the language model predict what the next word in the sequence will be.

"unigram model" => the previous words have no impact on the prediction.

# Language Model

- *Unigram language model*

- ▣ probability distribution over the words in a language
- ▣ generation of text according to the probability distribution

e.g., 5 words in total {"girl", "cat", "the", "cute", "touched"}

A possible language model: [0.2, 0.1, 0.35, 0.25, 0.1]

If we treat each document as a sequence of words, then the probabilities in the language model predict what the next word in the sequence will be.

**"unigram model" => the previous words have no impact on the prediction.**

**What is the probability to get the sequence "girl cat"? What about "girl touched"?**

# Language Model

- *Unigram language model*
  - ▣ probability distribution over the words in a language
  - ▣ generation of text according to the probability distribution
- *N-gram language model*
  - ▣ some applications (e.g., speech recognition) use bigram and trigram language models where probabilities depend on previous words
  - ▣ An n-gram model predicts a word based on the previous  $n - 1$  words

# Language Model

- *Unigram language model*
  - ▣ probability distribution over the words in a language
  - ▣ generation of text according to the probability distribution
- N-gram language model
  - ▣ some applications (e.g., speech recognition) use bigram and trigram language models where probabilities depend on previous words
  - ▣ An n-gram model predicts a word based on the previous  $n - 1$  words



# Language Model

- For search applications, we use language models to represent the topical content of a document.
- A *topic* in a document or query can be represented as a language model (i.e., a probability distribution over words)
  - ▣ i.e., words that tend to occur often when discussing a topic will have high probabilities in the corresponding language model

For example, if a document is about **fishing in Alaska**, we would expect to see words associated with **fishing and locations in Alaska** with high probabilities in the language model.

# Language Model (LM)

- For search applications, we use language models to represent the topical content of a document.
- A *topic* in a document or query can be represented as a language model
  - ▣ i.e., words that tend to occur often when discussing a topic will have high probabilities in the corresponding language model
- *Multinomial* distribution over words
  - ▣ text is modeled as a finite sequence of words, where there are  $t$  possible words at each point in the sequence
  - ▣ commonly used

# LMs for Retrieval

- In addition to representing documents as language models, we can also represent **the topic of the query** as a language model.

# LMs for Retrieval

- In addition to representing documents as language models, we can also represent **the topic of the query** as a language model.

The intuition is that the language model is **a representation of the topic that the information seeker had in mind when she was writing the query.**

# LMs for Retrieval

- In addition to representing documents as language models, we can also represent **the topic of the query** as a language model.
- 3 possibilities:
  - ▣ probability of generating the query text from a document language model
  - ▣ probability of generating the document text from a query language model
  - ▣ comparing the language models representing the query and document topics

Models of topical relevance

# Query-Likelihood Model

- Rank documents by the probability that the query could be generated by the document language model (i.e., **same topic**)

This is a model of topical relevance, in the sense that the probability of query generation is the measure of how likely it is that a document is about the same topic as the query

# Query-Likelihood Model

- Rank documents by the probability that the query could be generated by the document language model (i.e., same topic)
- Given a query  $Q$ , in general use  $P(D | Q)$  to rank documents
- Using Bayes' Rule
- Assuming prior is uniform, **unigram** model

$$p(D|Q) \stackrel{rank}{=} P(Q|D)P(D)$$

$$P(Q|D) = \prod_{i=1}^n P(q_i|D)$$

# Estimating Probabilities

- Obvious estimate for unigram probabilities is

$$P(q_i|D) = \frac{f_{q_i,D}}{|D|}$$

- *Maximum likelihood estimate*
  - ▣ makes the observed value of  $f_{q_i,D}$  most likely



# Estimating Probabilities

- Obvious estimate for unigram probabilities is

$$P(q_i|D) = \frac{f_{q_i,D}}{|D|}$$

- *Maximum likelihood estimate*

- ▣ makes the observed value of  $f_{q_i,D}$  most likely

- Problem: **If query words are missing from document, score will be zero**

- ▣ Missing 1 out of 4 query words same as missing 3 out of 4

# Smoothing

- Document texts are a *sample* from the language model
  - ▣ Missing words should not have zero probability of occurring
- *Smoothing* is a technique for estimating probabilities for missing (or unseen) words
  - ▣ lower (or *discount*) the probability estimates for words that are seen in the document text
  - ▣ assign that “left-over” probability to the estimates for the words that are not seen in the text

# Estimating Probabilities

- Estimate for unseen words is  $\alpha_D P(q_i | C)$ 
  - ▣  $P(q_i | C)$  is the probability for query word  $i$  in the collection language model for collection  $C$  (background probability)
  - ▣  $\alpha_D$  is a parameter
- Estimate for words that occur is
$$(1 - \alpha_D) P(q_i | D) + \alpha_D P(q_i | C)$$
- Different forms of estimation come from different  $\alpha_D$

# Jelinek-Mercer Smoothing

- $\alpha_D$  is a constant,  $\lambda$
- Gives estimate of

$$p(q_i|D) = (1 - \lambda) \frac{f_{q_i,D}}{|D|} + \lambda \frac{c_{q_i}}{|C|}$$

# Jelinek-Mercer Smoothing

- $\alpha_D$  is a constant,  $\lambda$

- Gives estimate of

$$p(q_i|D) = (1 - \lambda) \frac{f_{q_i,D}}{|D|} + \lambda \frac{c_{q_i}}{|C|}$$

- Ranking score

$$P(Q|D) = \prod_{i=1}^n \left( (1 - \lambda) \frac{f_{q_i,D}}{|D|} + \lambda \frac{c_{q_i}}{|C|} \right)$$

- Use logarithms for convenience

- ▣ accuracy problems with multiplying small numbers

$$\log P(Q|D) = \sum_{i=1}^n \log \left( (1 - \lambda) \frac{f_{q_i,D}}{|D|} + \lambda \frac{c_{q_i}}{|C|} \right)$$

# tf.idf Weight comes into play?

$$\begin{aligned}\log P(Q|D) &= \sum_{i=1}^n \log\left((1 - \lambda) \frac{f_{q_i,D}}{|D|} + \lambda \frac{c_{q_i}}{|C|}\right) \\ &= \sum_{i:f_{q_i,D}>0} \log\left((1 - \lambda) \frac{f_{q_i,D}}{|D|} + \lambda \frac{c_{q_i}}{|C|}\right) + \sum_{i:f_{q_i,D}=0} \log\left(\lambda \frac{c_{q_i}}{|C|}\right) \\ &= \sum_{i:f_{q_i,D}>0} \log \frac{\left((1 - \lambda) \frac{f_{q_i,D}}{|D|} + \lambda \frac{c_{q_i}}{|C|}\right)}{\lambda \frac{c_{q_i}}{|C|}} + \sum_{i=1}^n \log\left(\lambda \frac{c_{q_i}}{|C|}\right) \\ &\stackrel{\text{rank}}{=} \sum_{i:f_{q_i,D}>0} \log \left( \frac{\left((1 - \lambda) \frac{f_{q_i,D}}{|D|} + 1\right)}{\lambda \frac{c_{q_i}}{|C|}} \right)\end{aligned}$$

- proportional to the term frequency, inversely proportional to the collection frequency

# Dirichlet Smoothing

- $\alpha_D$  depends on document length

$$\alpha_D = \frac{\mu}{|D| + \mu}$$

- Gives probability estimation of

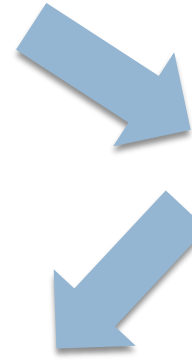
$$p(q_i | D) = \frac{f_{q_i, D} + \mu \frac{c_{q_i}}{|C|}}{|D| + \mu}$$

- and document score

$$\log P(Q | D) = \sum_{i=1}^n \log \frac{f_{q_i, D} + \mu \frac{c_{q_i}}{|C|}}{|D| + \mu}$$

General form:

$$(1 - \alpha_D) P(q_i | D) + \alpha_D P(q_i | C)$$



# Query Likelihood Example

- For the term “president”
  - $f_{qi,D} = 15, c_{qi} = 160,000$
- For the term “lincoln”
  - $f_{qi,D} = 25, c_{qi} = 2,400$
- number of word occurrences in the document is assumed to be 1,800
- number of word occurrences in the collection is  $10^9$
  
- $\mu = 2,000$



# Query Likelihood Example

$$\log P(Q|D) = \sum_{i=1}^n \log \frac{f_{q_i, D} + \mu \frac{c_{q_i}}{|C|}}{|D| + \mu}$$

$$\begin{aligned} QL(Q, D) &= \log \frac{15 + 2000 \times (1.6 \times 10^5 / 10^9)}{1800 + 2000} \\ &\quad + \log \frac{25 + 2000 \times (2400 / 10^9)}{1800 + 2000} \\ &= \log(15.32 / 3800) + \log(25.005 / 3800) \\ &= -5.51 + -5.02 = -10.53 \end{aligned}$$

- Negative number because summing logs of small numbers

# Query Likelihood Example

| Frequency of “president” | Frequency of “lincoln” | QL score |
|--------------------------|------------------------|----------|
| 15                       | 25                     | -10.53   |
| 15                       | 1                      | -13.75   |
| 15                       | 0                      | -19.05   |
| 1                        | 25                     | -12.99   |
| 0                        | 25                     | -14.40   |

- For the term “president”
  - $f_{q_i,D} = ?$ ,  $c_{q_i} = 160,000$
- For the term “lincoln”
  - $f_{q_i,D} = ?$ ,  $c_{q_i} = 2,400$

$$\log P(Q|D) = \sum_{i=1}^n \log \frac{f_{q_i,D} + \mu \frac{c_{q_i}}{|C|}}{|D| + \mu}$$

# Relevance Models

- *Relevance model* – language model representing information need
  - ▣ query and relevant documents are samples from this model (instead of calling it a query language, name it "relevance model")
- $P(D | R)$  - probability of generating the text in a document given a relevance model
  - ▣ *also called "document likelihood model"*
  - ▣ less effective than query likelihood due to difficulties comparing across documents of different lengths

# Pseudo-Relevance Feedback

- Estimate relevance model from query and top-ranked documents
- Rank documents by similarity of document model to relevance model
- *Kullback-Leibler divergence* (KL-divergence) is a well-known measure of the **difference** between two probability distributions

# KL-Divergence

- Given the *true* probability distribution  $P$  and another distribution  $Q$  that is an *approximation* to  $P$ ,

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$


- ▣ Use **negative KL-divergence** for ranking, and assume relevance model  $R$  is the true distribution (not symmetric),

$$\sum_{w \in V} P(w|R) \log P(w|D) - \sum_{w \in V} P(w|R) \log P(w|R)$$

# KL-Divergence

- Given a simple maximum likelihood estimate for  $P(w | R)$ , based on the frequency in the query text, ranking score is

$$\sum_{w \in V} \frac{f_{w,Q}}{|Q|} \log P(w|D)$$

- ▣ rank-equivalent to query likelihood score   $P(Q|D) = \prod_{i=1}^n P(q_i|D)$
- Query likelihood model is a special case of retrieval based on relevance model

# Estimating the Relevance Model

- Probability of pulling a word  $w$  out of the “bucket” representing the relevance model depends on the  $n$  query words we have just pulled out

$$P(w|R) \approx P(w|q_1 \dots q_n)$$

- By definition

$$P(w|R) \approx \frac{P(w, q_1 \dots q_n)}{P(q_1 \dots q_n)}$$

# Estimating the Relevance Model

- Joint probability is

$$P(w, q_1 \dots q_n) = \sum_{D \in \mathcal{C}} p(D) P(w, q_1 \dots q_n | D)$$

- Assume

$$P(w, q_1 \dots q_n | D) = P(w | D) \prod_{i=1}^n P(q_i | D)$$

- Gives

$$P(w, q_1 \dots q_n) = \sum_{D \in \mathcal{C}} P(D) P(w | D) \prod_{i=1}^n P(q_i | D)$$



# Estimating the Relevance Model

- $P(D)$  usually assumed to be uniform
- $P(w, q_1 \dots q_n)$  is simply a weighted average of the language model probabilities for  $w$  in a set of documents, where the **weights** are the query likelihood scores for those documents

$$P(w, q_1 \dots q_n) = \sum_{D \in \mathcal{C}} P(D) P(w|D) \underline{\prod_{i=1}^n P(q_i|D)}$$

- Formal model for pseudo-relevance feedback
  - ▣ query expansion technique

# Pseudo-Feedback Algorithm

$$P(w, q_1 \dots q_n) = \sum_{D \in \mathcal{C}} P(D) P(w|D) \prod_{i=1}^n P(q_i|D)$$

1. Rank documents using the query likelihood score for query  $Q$ .
2. Select some number of the top-ranked documents to be the set  $\mathcal{C}$ .
3. Calculate the relevance model probabilities  $P(w|R)$ .  $P(q_1 \dots q_n)$  is used as a normalizing constant and is calculated as

$$P(q_1 \dots q_n) = \sum_{w \in V} P(w, q_1 \dots q_n)$$

4. Rank documents again using the KL-divergence score

$$P(w|R) \approx \frac{P(w, q_1 \dots q_n)}{P(q_1 \dots q_n)}$$

$$\sum_w P(w|R) \log P(w|D)$$

# Example (most probable words in $p(w | R)$ based on Top 10 Docs)

| <i>president lincoln</i> | <i>abraham lincoln</i> | <i>fishing</i> | <i>tropical fish</i> |
|--------------------------|------------------------|----------------|----------------------|
| lincoln                  | lincoln                | fish           | fish                 |
| president                | america                | farm           | tropic               |
| room                     | president              | salmon         | japan                |
| bedroom                  | faith                  | new            | aquarium             |
| house                    | guest                  | wild           | water                |
| white                    | abraham                | water          | species              |
| america                  | new                    | caught         | aquatic              |
| guest                    | room                   | catch          | fair                 |
| serve                    | christian              | tag            | china                |
| bed                      | history                | time           | coral                |
| washington               | public                 | eat            | source               |
| old                      | bedroom                | raise          | tank                 |
| office                   | war                    | city           | reef                 |
| war                      | politics               | people         | animal               |
| long                     | old                    | fishermen      | tarpon               |
| abraham                  | national               | boat           | fishery              |

# Example (most probable words in $p(w | R)$ based on Top 10 Docs)

| <i>president lincoln</i> | <i>abraham lincoln</i> | <i>fishing</i> | <i>tropical fish</i> |
|--------------------------|------------------------|----------------|----------------------|
| lincoln                  | lincoln                | fish           | fish                 |
| president                | president              | water          | tropic               |
| america                  | america                | catch          | water                |
| new                      | abraham                | reef           | storm                |
| national                 | war                    | fishermen      | species              |
| great                    | man                    | river          | boat                 |
| white                    | civil                  | new            | sea                  |
| war                      | new                    | year           | river                |
| washington               | history                | time           | country              |
| clinton                  | two                    | bass           | tuna                 |
| house                    | room                   | boat           | world                |
| history                  | booth                  | world          | million              |
| time                     | time                   | farm           | state                |
| center                   | politics               | angle          | time                 |
| kennedy                  | public                 | fly            | japan                |
| room                     | guest                  | trout          | mile                 |

# Combining Evidence

- Effective retrieval requires the combination of many pieces of evidence about a document's potential relevance
  - ▣ have focused on simple word-based evidence
  - ▣ many other types of evidence
    - structure, PageRank, metadata, even scores from different models
- *Inference network* model is one approach to combining evidence
  - ▣ uses Bayesian network formalism

107

## Backup slides

# Galago Query Language

- A document is viewed as a sequence of text that may contain arbitrary tags
- A single *context* is generated for each unique tag name
- An *extent* is a sequence of text that appears within a single begin/end tag pair of the same type as the context

# Galago Query Language

```
<html>
<head>
<title>Department Descriptions</title>
</head>
<body>
The following list describes ...
<h1>Agriculture</h1> ...
<h1>Chemistry</h1> ...
<h1>Computer Science</h1> ...
<h1>Electrical Engineering</h1> ...
</body>
</html>
```

title context:

```
<title>Department Descriptions</title>
```

h1 context:

```
<h1>Agriculture</h1>
<h1>Chemistry</h1> ...
<h1>Computer Science</h1> ...
<h1>Electrical Engineering</h1> ...
```

body context:

```
<body> The following list describes ...
<h1>Agriculture</h1> ...
<h1>Chemistry</h1> ...
<h1>Computer Science</h1> ...
<h1>Electrical Engineering</h1> ...
</body>
```



# Galago Query Language

*Simple terms:*

term – term that will be normalized and stemmed.

"term" – term is not normalized or stemmed.

*Examples:*

presidents

"NASA"

# Galago Query Language

## *Proximity terms:*

`#od:N( ... )` – ordered window – terms must appear ordered, with at most N-1 terms between each.

`#od( ... )` – unlimited ordered window – all terms must appear ordered anywhere within current context.

`#uw:N( ... )` – unordered window – all terms must appear within a window of length N in any order.

`#uw( ... )` – unlimited unordered window – all terms must appear within current context in any order.

## *Examples:*

`#od:1(white house)` – matches “white house” as an exact phrase.

`#od:2(white house)` – matches “white \* house” (where \* is any word or null).

`#uw:2(white house)` – matches “white house” and “house white”.

# Galago Query Language

*Synonyms:*

`#syn( ... )`

`#wsyn( ... )`

*Examples:*

`#syn(dog canine)` – simple synonym based on two terms.

`#syn( #od:1(united states) #od:1(united states of america) )` – creates a synonym from two proximity terms.

`#wsyn( 1.0 donald 0.8 don 0.5 donnie )` – weighted synonym indicating relative importance of terms.

# Galago Query Language

*Anonymous terms:*

`#any:.`() – used to match extent types

*Examples:*

`#any:person()` – matches any occurrence of a person extent.

`#od:1(lincoln died in #any:date())` – matches exact phrases of the form: “lincoln died in <date>... </date>”.

# Galago Query Language

## *Context restriction and evaluation:*

`expression.C1,...,CN` – matches when the expression appears in all contexts C1 through CN.

`expression.(C1,...,CN)` – evaluates the expression using the language model defined by the concatenation of contexts C1...CN within the document.

### *Examples:*

`dog.title` – matches the term “dog” appearing in a title extent.

`#uw(smith jones).author` – matches when the two names “smith” and “jones” appear in an author extent.

`dog.(title)` – evaluates the term based on the title language model for the document.

`#od:1(abraham lincoln).person.(header)` – builds a language model from all of the “header” text in the document and evaluates `#od:1(abraham lincoln).person` in that context (i.e., matches only the exact phrase appearing within a person extent within the header context).

# Galago Query Language

## *Belief operators:*

`#combine(...)` – this operator is a normalized version of the  $bel_{and}(q)$  operator in the inference network model. See the discussion below for more details.

`#weight(...)` – this is a normalized version of the  $bel_{wand}(q)$  operator.

`#filter(...)` – this operator is similar to `#combine`, but with the difference that the document must contain at least one instance of all terms (simple, proximity, synonym, etc.). The evaluation of nested belief operators is not changed.

# Galago Query Language

*Examples:*

`#combine( #syn(dog canine) training )` – rank by two terms, one of which is a synonym.

`#combine( biography #syn(#od:1(president lincoln) #od:1(abraham lincoln)) )` – rank using two terms, one of which is a synonym of “president lincoln” and “abraham lincoln”.

`#weight( 1.0 #od:1(civil war) 3.0 lincoln 2.0 speech )` – rank using three terms, and weight the term “lincoln” as most important, followed by “speech”, then “civil war”.

`#filter( aquarium #combine(tropical fish) )` – consider only those documents containing the word “aquarium” and “tropical” or “fish”, and rank them according to the query `#combine(aquarium #combine(tropical fish))`.

`#filter( #od:1(john smith).author) #weight( 2.0 europe 1.0 travel )` – rank documents about “europe” or “travel” that have “John Smith” in the author context.

# Web Search

- Most important, but not only, search application
- Major differences to TREC news
  - ▣ Size of collection
  - ▣ Connections between documents
  - ▣ Range of document types
  - ▣ Importance of spam
  - ▣ Volume of queries
  - ▣ Range of query types



# Search Taxonomy

- *Informational*
  - ▣ Finding information about some topic which may be on one or more web pages
  - ▣ Topical search
- *Navigational*
  - ▣ finding a particular web page that the user has either seen before or is assumed to exist
- *Transactional*
  - ▣ finding a site where a task such as shopping or downloading music can be performed

# Web Search

- For effective navigational and transactional search, need to combine features that reflect *user relevance*
- Commercial web search engines combine evidence from *hundreds* of features to generate a ranking score for a web page
  - ▣ page content, page metadata, anchor text, links (e.g., PageRank), and user behavior (click logs)
  - ▣ page metadata – e.g., “age”, how often it is updated, the URL of the page, the domain name of its site, and the amount of text content

# Search Engine Optimization

- *SEO*: understanding the relative importance of features used in search and how they can be manipulated to obtain better search rankings for a web page
  - ▣ e.g., improve the text used in the title tag, improve the text in heading tags, make sure that the domain name and URL contain important keywords, and try to improve the anchor text and link structure
  - ▣ Some of these techniques are regarded as not appropriate by search engine companies

# Web Search

- In TREC evaluations, most effective features for navigational search are:
  - ▣ text in the title, body, and heading (h1, h2, h3, and h4) parts of the document, the anchor text of all links pointing to the document, the PageRank number, and the inlink count
- Given size of Web, many pages will contain all query terms
  - ▣ Ranking algorithm focuses on discriminating between these pages
  - ▣ Word proximity is important

# Term Proximity

- Many models have been developed
- N-grams are commonly used in commercial web search
- *Dependence model* based on inference net has been effective in TREC - e.g.

```
#weight(  
  0.8 #combine(embryonic stem cells)  
  0.1 #combine( #od:1(stem cells) #od:1(embryonic stem  
               #od:1(embryonic stem cells))  
  0.1 #combine( #uw:8(stem cells) #uw:8(embryonic cells)  
               #uw:8(embryonic stem) #uw:12(embryonic stem cells)))
```

# Example Web Query

```
#weight(  
  0.1 #weight( 0.6 #prior(pagerank) 0.4 #prior(inlinks))  
  1.0 #weight(  
    0.9 #combine(  
      #weight( 1.0 pet.(anchor) 1.0 pet.(title)  
              3.0 pet.(body) 1.0 pet.(heading))  
      #weight( 1.0 therapy.(anchor) 1.0 therapy.(title)  
              3.0 therapy.(body) 1.0 therapy.(heading)))  
    0.1 #weight(  
      1.0 #od:1(pet therapy).(anchor) 1.0 #od:1(pet therapy).(title)  
      3.0 #od:1(pet therapy).(body) 1.0 #od:1(pet therapy).(heading))  
    0.1 #weight(  
      1.0 #uw:8(pet therapy).(anchor) 1.0 #uw:8(pet therapy).(title)  
      3.0 #uw:8(pet therapy).(body) 1.0 #uw:8(pet therapy).(heading)))  
  )  
)
```

# Machine Learning and IR

- Considerable interaction between these fields
  - ▣ Rocchio algorithm (60s) is a simple learning approach
  - ▣ 80s, 90s: learning ranking algorithms based on user feedback
  - ▣ 2000s: text categorization
- Limited by amount of training data
- Web query logs have generated new wave of research
  - ▣ e.g., “Learning to Rank”

# Generative vs. Discriminative

- All of the probabilistic retrieval models presented so far fall into the category of *generative models*
  - ▣ A generative model assumes that documents were generated from some underlying model (in this case, usually a multinomial distribution) and uses training data to estimate the parameters of the model
  - ▣ probability of belonging to a class (i.e. the relevant documents for a query) is then estimated using Bayes' Rule and the document model



# Generative vs. Discriminative

- A *discriminative* model estimates the probability of belonging to a class directly from the observed features of the document based on the training data
- Generative models perform well with low numbers of training examples
- Discriminative models usually have the advantage given enough training data
  - ▣ Can also easily incorporate many features

# Discriminative Models for IR

- Discriminative models can be trained using explicit relevance judgments or click data in query logs
  - ▣ Click data is much cheaper, more noisy
  - ▣ e.g. Ranking Support Vector Machine (SVM) takes as input *partial rank* information for queries
    - partial information about which documents should be ranked higher than others

# Ranking SVM

- Training data is

$$(q_1, r_1), (q_2, r_2), \dots, (q_n, r_n)$$

- $r$  is partial rank information

- if document  $d_a$  should be ranked higher than  $d_b$ , then  $(d_a, d_b) \in r_i$

- partial rank information comes from relevance

judgments (allows multiple levels of relevance) or click data

- e.g.,  $d_1, d_2$  and  $d_3$  are the documents in the first, second and third rank of the search output, only  $d_3$  clicked on  $\rightarrow (d_3, d_1)$  and  $(d_3, d_2)$  will be in desired ranking for this query

# Ranking SVM

- Learning a linear ranking function  $\vec{w} \cdot \vec{d}_a$ 
  - ▣ where  $w$  is a weight vector that is adjusted by learning
  - ▣  $d_a$  is the vector representation of the features of document
  - ▣ *non-linear* functions also possible
- Weights represent importance of features
  - ▣ learned using training data
  - ▣ e.g.,

$$\vec{w} \cdot \vec{d} = (2, 1, 2) \cdot (2, 4, 1) = 2 \cdot 2 + 1 \cdot 4 + 2 \cdot 1 = 10$$

# Ranking SVM

- Learn  $w$  that satisfies as many of the following conditions as possible:

$$\forall (d_i, d_j) \in r_1 \quad : \quad \vec{w} \cdot \vec{d}_i > \vec{w} \cdot \vec{d}_j$$

...

- Can be formula  $\forall (d_i, d_j) \in r_n \quad : \quad \vec{w} \cdot \vec{d}_i > \vec{w} \cdot \vec{d}_j$  problem

# Ranking SVM

$$\text{minimize : } \frac{1}{2} \vec{w} \cdot \vec{w} + C \sum \xi_{i,j,k}$$

*subject to :*

$$\forall (d_i, d_j) \in r_1 \quad : \quad \vec{w} \cdot \vec{d}_i > \vec{w} \cdot \vec{d}_j + 1 - \xi_{i,j,1}$$

...

$$\forall (d_i, d_j) \in r_n \quad : \quad \vec{w} \cdot \vec{d}_i > \vec{w} \cdot \vec{d}_j + 1 - \xi_{i,j,n}$$

$$\forall i \forall j \forall k : \xi_{i,j,k} \geq 0$$

- ▣  $\xi$ , known as a slack variable, allows for misclassification of difficult or noisy training examples, and  $C$  is a parameter that is used to prevent overfitting

# Ranking SVM

- Software available to do optimization
- Each pair of documents in our training data can be represented by the vector:

$$(\vec{d}_i - \vec{d}_j)$$

- Score for this pair is:

$$\vec{w} \cdot (\vec{d}_i - \vec{d}_j)$$

- SVM classifier will find a  $w$  that makes the smallest score as large as possible
  - ▣ make the differences in scores as large as possible for the pairs of documents that are hardest to rank

# Topic Models

- Improved representations of documents
  - ▣ can also be viewed as improved smoothing techniques
  - ▣ improve estimates for words that are related to the topic(s) of the document
    - instead of just using background probabilities
- Approaches
  - ▣ *Latent* Semantic Indexing (LSI)
  - ▣ Probabilistic *Latent* Semantic Indexing (pLSI)
  - ▣ *Latent* Dirichlet Allocation (LDA)



# LDA

- Model document as being generated from a *mixture* of topics
  1. For each document  $D$ , pick a multinomial distribution  $\theta_D$  from a Dirichlet distribution with parameter  $\alpha$ ,
  2. For each word position in document  $D$ ,
    - (a) pick a topic  $z$  from the multinomial distribution  $\theta_D$ ,
    - (b) Choose a word  $w$  from  $P(w|z, \beta)$ , a multinomial probability conditioned on the topic  $z$  with parameter  $\beta$ .

# LDA

- Gives language model probabilities

$$P_{lda}(w|D) = P(w|\theta_D, \beta) = \sum_z P(w|z, \beta)P(z|\theta_D)$$

- Used to smooth the document representation by mixing them with the query likelihood probability as follows:

$$P(w|D) = \lambda \left( \frac{f_{w,D} + \mu \frac{c_w}{|C|}}{|D| + \mu} \right) + (1 - \lambda)P_{lda}(w|D)$$

# LDA

- If the LDA probabilities are used directly as the document representation, the effectiveness will be significantly reduced because the features are *too smoothed*
  - ▣ e.g., in typical TREC experiment, only 400 topics used for the *entire* collection
  - ▣ generating LDA topics is expensive
- When used for smoothing, effectiveness is improved

# LDA Example

## ▣ Top words from 4 LDA topics from TREC news

| <i>Arts</i> | <i>Budgets</i> | <i>Children</i> | <i>Education</i> |
|-------------|----------------|-----------------|------------------|
| new         | million        | children        | school           |
| film        | tax            | women           | students         |
| show        | program        | people          | schools          |
| music       | budget         | child           | education        |
| movie       | billion        | years           | teachers         |
| play        | federal        | families        | high             |
| musical     | year           | work            | public           |
| best        | spending       | parents         | teacher          |
| actor       | new            | says            | bennett          |
| first       | state          | family          | manigat          |
| york        | plan           | welfare         | namphy           |
| opera       | money          | men             | state            |
| theater     | programs       | percent         | president        |
| actress     | government     | care            | elementary       |
| love        | congress       | life            | haiti            |

# Summary

- Best retrieval model depends on application and data available
- Evaluation corpus (or test collection), training data, and user data are all critical resources
- Open source search engines can be used to find effective ranking algorithms
  - ▣ Galago query language makes this particularly easy
- Language resources (e.g., thesaurus) can make a big difference